Does Category A Anchor Text Improve Category B Results?

Leonid Boytsov and Anna Belova
leo@boytsov.info, anna@belova.org

We merged results obtained from the Category B index with results obtained from the index built over complete (Category A) anchor text. However, we were unable to improve over Category B results in either the ad hoc or the diversity task.

1. INTRODUCTION

Associating anchor text with pages, to which links are pointing, is a well-known approach to improve retrieval quality. It was used in the first version of Google [Brin and Page 1998]. On one hand, using the anchor text alone allows one to obtain a system with decent performance [Anh and Moffat 2010; Hiemstra and Hauff 2010]. We also know that the anchor text is a strong relevance signal from our own experiments in TREC 2011 [Boytsov and Belova 2011]. On the other hand, the size of the anchor text is much smaller than size of the text for a full collection. Thus, enriching the Category B index (built over 50M documents) with the Category A anchor text index (built over 370M short documents), seemed to be an appealing method of improving performance at little cost.

2. EXPERIMENTS

2.1 Setup

We used two retrieval engines. One was a system developed for TREC 2011, which included an index for the Category B subset (50M documents). It explicitly indexed posting lists of close word pairs (where at least one word was frequent) and had a large index of 513 Gb. The detailed description of this type of index is given in our 2010 and 2011 reports [Boytsov and Belova 2010; 2011]. There are more than 20 relevance features combined in a semi-linear formula. In TREC 2011, we showed that this system was a strong benchmark: See the run srchvrs11b (Table 2) in the overview paper by Clarke et al. [2011].

In addition, we built a similar index over the Category A anchor text (Category A anchor text was compiled by Hiemstra and Hauff [2010]). Unlike the Category B index, it employed only one text field: anchor text. The anchor text index relied on the SpamRank, but not on the PageRank. The number of documents was about 370M. However, each document was small and the size of the index was only 212 Gb.

2.2 Results

We tried several approaches to combine scores from two retrieval systems: a linear combination of scores with different dictionaries, a linear combination of scores with the shared dictionary, and a round-robin method. In the approach with the shared

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Does Category A Anchor Text Improve Category B Results?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University ,Language Technologies Institute,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**3**	

dictionary, we used IDF values only from the Category B dictionary. None of the approaches allowed us to achieve higher performance scores on training data in the ad hoc task. However, we obtained slightly higher values of the diversity metric α-nDCG@20.

Overall, we submitted three runs srchvrs12c10, srchvrs12c09, and srchvrs12c00, where anchor text scores were summed up with with Category B scores. Prior to aggregating, anchor text scores were multiplied by the scaling coefficients 1, 0.9, and 0 respectively. The last run (srchvrs12c00) represents a "pure" Category B run.

Table I: Comparing performance of runs based on Category A anchor text against performance of Category B runs (for different years).

year	2010	2011	2012
anchor text	0.056	0.084	0.079
Category B run	0.106	0.137	0.307

Scores are computed using ERR@20

It turned out that all three runs had almost identical diversity scores ERR-IA@20, which were approximately equal to 0.38. The ad hoc scores were very similar as well: for example, ERR@20 was approximately equal to 0.305. We see that both srchvrs12c09 and srchvrs12c10 improved in MAP over the pure Category B run srchvrs12c00 (this was a statistically significant improvement that "survived" the Holm-Bonferroni correction). Yet, these improvements were small: 2.3% and 5%, respectively.

We also compared performance of runs that relied solely on Category A anchor text with performance of Category B runs (same algorithm as for srchvrs12c00). According to Table I, in 2010-2011 the values of ERR@20 for anchor text runs were only slightly higher than 1/2 of ERR@20 scores for the respective Category B runs. In 2012, however, the anchor text run had almost 4x weaker performance compared to the Category B run. This may partially explain the fact that combining anchor text runs and Category B runs did not lead to noticeable improvement in performance.

Finally, we evaluated an effect of not using SpamRank in 2010, 2011, and 2012. To do this, we set the SpamRank factor to 1 (it is included multiplicatively). We found that 2010 was the only year in which the SpamRank improved ERR@20 scores of our method. This is in contrast with our 2010 observation that SpamRanks can improve performance scores by a large margin [Boytsov and Belova 2010]. Perhaps, a more advanced system, which, among other factors, includes anchor text, is more robust to spam. It may also indicate that embedding a good relevance feature into an already strong baseline does not necessarily lead to a performance boost [Armstrong et al. 2009].

3. CONCLUSIONS

We merged results obtained from the Category B index with results obtained from the index built over complete (Category A) anchor text. Yet, this approach did not lead to a significant improvement in performance. We hypothesize that simple merging approaches (such as linear combinations or round-robin) do not work well

if one of the systems has a much lower performance than the other.

REFERENCES

Anh, V. N. and Moffat, A. 2010. The role of anchor text in clueweb09 retrieval. In *TREC*.

Armstrong, T. G., Moffat, A., Webber, W., and Zobel, J. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management*. CIKM '09. ACM, New York, NY, USA, 601–610.

Boytsov, L. and Belova, A. 2010. Lessons learned from indexing close word pairs. In *TREC-19: Proceedings of the Nineteenth Text REtrieval Conference*.

Boytsov, L. and Belova, A. 2011. Lessons learned from indexing close word pairs. In *TREC-20: Proceedings of the Nineteenth Text REtrieval Conference*.

Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems 30,* 17, 107 – 117. Proceedings of the Seventh International World Wide Web Conference.

Clarke, C. L. A., Craswell, N., Soboro, I., and Voorhees, E. M. 2011. Overview of the trec 2011 web track. In *TREC-20: Proceedings of the Nineteenth Text REtrieval Conference*.

Hiemstra, D. and Hauff, C. 2010. Mirex: Mapreduce information retrieval experiments. Technical Report TR-CTIT-10-15, Centre for Telematics and Information Technology University of Twente, Enschede. April.